2

Naval Research Laboratory

Washington, DC 20375-5000

NRL/FR/5530—92-9379

Intelligibility and Acceptability Testing for Speech Technology

ASTRID SCHMIDT-NIELSEN

Human Computer Interaction Laboratory
Information Technology Division

May 22, 1992

Approved for public release; distribution unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
May 22, 1992

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Intelligibility and Acceptability Testing
for Speech Technology

5. FUNDING NUMBERS
PE - 33904N
WU - DN619-111

6. AUTHOR(S)
Astrid Schmidt-Nielsen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Research Laboratory
Washington, DC 20375-5000

8. PERFORMING ORGANIZATION REPORT NUMBER
NRL/FR/5530—92-9379

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Space and Naval Warfare Systems Command
Washington, DC 20363-5100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The evaluation of speech intelligibility and acceptability is an important aspect of the use, development, and selection of voice communication devices—telephone systems, digital voice systems, speech synthesis by rule, speech in noise, and the effects of noise stripping. Standard test procedures can provide highly reliable measures of speech intelligibility, and subjective acceptability tests can be used to evaluate voice quality. These tests are often highly correlated with other measures of communication performance and can be used to predict performance in many situations. However, when the speech signal is severely degraded or highly processed, a more complete evaluation of speech quality is needed—one that takes into account the many different sources of information that contribute to how we understand speech.

14. SUBJECT TERMS
Intelligibility testing Voice quality Speech quality testing
Acceptability testing Speech evaluation

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

CONTENTS

INTELLIGIBILITY AND ACCEPTABILITY TESTING FOR SPEECH TECHNOLOGY

INTRODUCTION

The need for evaluation occurs at many stages of speech technology development—during development to determine whether improvement has occurred, in manufacturing to determine if specifications are met, and in selection to compare and choose among competing equipment or techniques. To evaluate the performance of a speech communication, processing, or synthesis system, some form of intelligibility or acceptability testing involving human listeners is usually conducted. The tests that are used can vary considerably in sophistication and reliability. In intelligibility testing, one or more listeners perform a task in which they listen to the transmitted or synthesized speech and report what they hear. Depending on the specific test, the listener's task may be to write down a sentence, word, or sound or to select the response that most closely matches what was heard from two or more alternatives. The intelligibility score is then based on the percentage of correct responses. In contrast to measures of speech intelligibility, which can be objectively scored, evaluations of speech acceptability are based on subjective listener judgments. Subjects listen to speech samples and rate the quality of the speech by using either a numerical scale or verbal labels, which can later be converted to numbers.

The problem of conducting intelligibility and acceptability tests to evaluate voice communication systems is especially difficult because the listeners and speakers that are used to evaluate the speech can vary considerably among individuals and over time, whereas the performance of the equipment itself is highly stable from one time to another and among different units of the same model. This report discusses intelligibility and acceptability test methods. The use of physical measures of the speech signal as possible indices of speech intelligibility or acceptability is also briefly considered. An overview of the testing process and a discussion of the factors that contribute to speech intelligibility and acceptability precede a review of intelligibility and acceptability test methods. Finally, experimental relations among different tests and considerations in selecting test methods will be discussed, and some general recommendations will be made.

The general concept of intelligibility refers to how well the speech can be comprehended or understood. Speech that is more intelligible is easier to understand. This leads more or less directly to the notion of measuring intelligibility by counting the number of words or speech sounds that are correctly understood. In real life, however, factors other than the fidelity of the acoustic signal also contribute to how well the speech is understood (e.g., listener expectations and subject matter). It is important to remember that the score obtained on an intelligibility test is only a predictor or estimate of what we really want to know and not an end in itself. The goal of testing is not merely to obtain high scores but to develop usable systems.

Intelligibility testing is a compromise that requires trade-offs among conflicting goals and sometimes incompatible test requirements. One of the most important goals is to determine the usability of a communication system for a given application. Potential users of the equipment want to know how well

Manuscript approved January 1, 1992.

the system will perform in operational environments with realistic vocabularies. It is also highly desirable to be able to compare the results of different test conditions with one another. Decision makers who use scores for selection prefer to deal with exact numbers; they want to know that the score is 92 and do not want to be told that next time the score may be 87 even if the system would still have the same rank ordering as before. For the purpose of writing procurement or manufacturing specifications, a specific criterion is needed—i.e., that the score meets or exceeds the specified value. Practical considerations of time, cost, and human resources are also important in determining the kind of testing that can be carried out. Realistic field tests are time consuming and expensive, and it is often difficult to quantify the results. Field tests are also highly specific to a particular situation and do not generalize well to other situations, nor do they allow for meaningful comparisons of test scores with the results of tests made under other conditions. The result is that intelligibility testing usually involves relatively simple listening tasks, so that the test may be given and scored quickly and inexpensively, and so that the test does not end up evaluating extraneous factors rather than the performance of the speech system. These intelligibility test tasks differ from the tasks of the actual use of a system, which might involve, for example, specialized vocabulary and grammar, rapid responses in emergencies, or fatigue through extended use. Unfortunately, the test materials that give the most repeatable results, like rhyme tests or nonsense syllable tests, are often the least realistic, while the most realistic speech materials—sentences or conversations—tend to produce the least repeatable results. If the correlations among tests that use different types of speech materials were entirely consistent, it would be possible to predict performance with more realistic vocabularies and speech materials from rhyme test scores. However, as will be seen later in this report, generalizing from one type of test to another is questionable, especially when comparing very different speech systems or degradations.

In selecting and using speech intelligibility and acceptability tests and in interpreting the test scores, it is important to understand how the test tasks relate to actual tasks in order to make intelligent decisions about when and how test results can be extrapolated to performance in the real world. The most obvious consequences of poor speech intelligibility are mistakes in understanding the spoken message such as misperceptions, confusions among words, or words that are missed altogether; this of course is the reason for intelligibility tests. Poor speech quality can have consequences, even when all of the words are correctly understood. When more effort is required to understand the speech, it may be harder to integrate and store the information in memory (Luce, Feustel, and Pisoni, 1983). In high workload or multiple task situations, the added effort of listening to degraded speech can lead to poorer performance on concurrent tasks (Schmidt-Nielsen, Kallman, and Meijer, 1990). Many of these effects are difficult to track or quantify because the human being is an extremely flexible information processor who develops new strategies or uses more effort to compensate for the deficiencies in the system and thus manages to perform the required task in spite of poor speech quality. The long-term effects of listening to degraded speech over extended periods of time, such as fatigue and stress, are even more difficult to document, but are probably nonetheless real. Standard intelligibility tests can provide stable and repeatable scores that make it possible to compare results across different conditions, but it is often useful to supplement standard test results with other tests and experiments to evaluate other aspects of system performance.

OVERVIEW OF THE TESTING PROCESS

Figure 1 shows the principal elements in the testing of voice systems. These are also the basic elements of a voice transmission system. For simplicity, a number of links have been left out of this simplified system, for example an input device such as a microphone and an output device such as a loudspeaker or headphones. Actual communication systems include two-way transmission, and the voice processor/coder at each end has both analysis and synthesis components. The elements in the diagram may be subject to considerable elaboration, and all of the elements need not be present for any given test situation. For example, the testing of speech synthesizers includes only the right-hand side of the diagram, and, in some cases, tests of the transmission channel may be omitted.

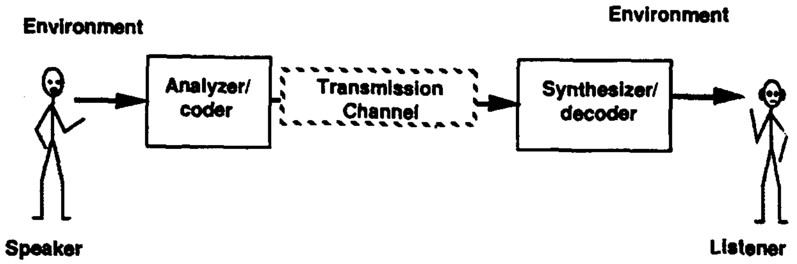

Fig. 1 — Simplified diagram of the elements needed in a communication system and for intelligibility testing. Actual communication systems include two-way trnasmission, and the voice processor at each end has both analysis and synthesis components.

Characteristics of each of the elements depicted in Fig. 1 affect the outcome of the testing process and can also interact with the other elements in determining the final score. Some of these characteristics (e.g., background noise) may be of specific interest in the evaluation process and may be systematically varied in a test series. Others (e.g., listener differences) may only contribute random variability to the test scores and should be carefully controlled. In deciding how to conduct a series of tests, the tester should have a clear idea of which aspects are of interest in the evaluation process and which ones need to be controlled.

The speaker. Different speakers have different voice characteristics that affect the performance and intelligibility of the voice system. It is well recognized that some voices, especially female voices, can cause problems for narrowband digital voice systems such as linear predictive coding (LPC) algorithms. A speaker whose voice performs well in one environment may not be the best in another.

The speaker environment. Voice communication systems are often used in environments where background noise is present. This may vary from relatively benign environments like an office to severe environments such as a helicopter or tank. Even though people generally tend to speak more loudly when background noise is present; some voices are considerably more intelligible in noise than others. The background noise environment not only affects the voice characteristics of the speaker by causing changes in amplitude, pitch, duration, and formant frequencies (e.g., Summers, Pisoni, Bernacki, Pedlow, and Stokes, 1988), but when noise enters the microphone, the performance of the voice system may also be degraded. Some communication systems degrade considerably in the presence of background noise, while others are quite robust. Systems designed primarily to handle speech may be particularly susceptible to degradation from nonspeech sounds.

The voice processor. Digital voice transmission systems, wideband and narrowband, consist of a speech analyzer at the input, which analyses and codes the speech signal, and a synthesizer at the output to reconstruct the speech signal and transform it back to analog voice output. Noise reduction by various techniques is a form of speech processing that may be used either at the input of a communication device before the speech is coded and transmitted or at the output after noisy speech is received. Speech synthesis for computer voice output uses rules or stored segments to generate the voice instead of the human speaker.

The transmission channel. When a communication system is tested in back-to-back mode, and the output of the analysis portion goes directly into the synthesis portion, the only degradation is due to the voice processor. This is the best performance that can be expected. In actual use, the transmission channel for voice communication systems is another possible source of degradation. Telephone links may suffer from echo or crosstalk. Radio transmissions may be subject to various forms of interference, natural or man-made. Digital voice processors are susceptible to bit errors, random or burst, from various sources of interference. Digital voice transmissions may sometime involve tandems, i.e., more than one digital processing link, with the speech being converted back to analog form before transmission through the next link. Combinations of degradations may interact with one another to cause even more severe degradation.

The listener. Listeners vary in their ability to make speech discriminations. When listeners with normal hearing are used, listener variability is generally smaller than speaker variability for most intelligibility tests (Voiers, 1982). Listener variability tends to be greater than speaker variability for acceptability tests (Voiers, personal communication). In more complex speech tasks that involve sentence or dialogue comprehension, other listener's skills such as the attention span or language ability, can also affect the results.

The listening environment. Like the speaking environment, the listening environment may have more or less severe background noise that can affect speech intelligibility. However, the listener can often compensate by increasing the volume of the speech.

Tape recording is frequently used at various stages of the testing process for standardization and control of the test materials as well as for portability and reproduceability. The speakers used in the intelligibility test are recorded while reading the test materials and the recorded materials are processed through the voice equipment. The output can also be recorded for later use in testing and played back to the listeners. Background noises of various types may be recorded and their levels noted, so they can later be played back at the same levels to simulate different speaking or listening environments.

FACTORS THAT INFLUENCE SPEECH INTELLIGIBILITY AND ACCEPTABILITY

When speech is used for communication, information that can be used to decode and understand the speech sounds is available on many different levels. Human listeners are versatile and efficient processors of information with remarkable capabilities for understanding even a very degraded speech signal. They will use every available source of information to perform the speech task. Several types of information are available to the listener in a typical communication situation. There is the acoustic-phonetic information in the speech signal itself as well as contextual information from the surrounding circumstances. The information carried by the speech signal includes segmental information—the acoustic-phonetic cues for consonant and vowel identity, and suprasegmental or prosodic information—the intonation, timing, and intensity cues that carry information about word stress and word and sentence structure. Words spoken in isolation carry more acoustic detail relating to the phonetic structure of the word while prosody and context assume greater importance in understanding connected speech. The words in spoken sentences are highly intelligible within the sentence context, but individual words extracted from such speech are poorly recognized (Pollack and Pickett, 1964). Contextual information can include the grammatical and semantic constraints imposed by sentence structure as well as specific situational knowledge and general knowledge that influence the listener's expectations. If little information is available from the context, as for example in trying to identify nonsense syllables, the listener must rely almost exclusively on acoustic-phonetic information in the signal for correct identification. Conversely, if more information is available from the context, the listener needs less accurate acoustic information to correctly identify the speech. Different types of tests present different levels of information to the listener and measure different aspects of the comprehension process. In this section we consider some of the types of information that people may use in the communication process, how the distortion of this information by a voice processing system might affect the intelligibility or judged acceptability of a system, and the extent to which it is measured by existing tests.

Segmental information. The acoustic-phonetic information that allows us to identify individual speech sounds needs to be correctly transmitted for the speech to be intelligible. In analysis-synthesis voice transmission systems, especially low data rate systems, the acoustic cues needed to correctly identify the speech sounds are not always correctly reproduced, and this reduces intelligibility. Speech synthesis by rule does not always produce the correct sounds and the effects of coarticulation (the influence of adjacent speech sounds on one another) may also be incorrectly reproduced in some contexts, so that phoneme intelligibility may vary considerably depending on where the sound occurs in the word. Intelligibility tests explicitly test phoneme intelligibility, but the standard tests include discriminations only in initial and final positions in single syllable words.

Suprasegmental information. Prosodic information (the variation in pitch, intensity, and timing across segments) conveys information about stress levels and word and sentence structure, as well as the more subtle nuances of meaning and emotion. In spoken English, the intonation is often the only cue that distinguishes a statement from a question. (Compare "He did?" with rising intonation and "He did." with falling intonation.) Some low data rate and extreme low data rate digital voice transmission systems may have problems with pitch and intensity, especially if rapid changes are present, but for most voice transmission systems, pitch and intensity changes as well as segmental timing tend to be reasonably accurately reproduced, so testing prosody explicitly is not usually a problem. In speech synthesis by rule, intonation and timing rules that follow the constraints of the spoken language must be explicitly built into the system, and the effect of prosody on intelligibility and naturalness needs to be evaluated. The extent to which prosodic information is correctly conveyed is usually not explicitly tested in intelligibility tests, although tests using sentences or paragraphs may implicitly test the goodness of the prosody. Incorrect or distorted prosody also tends to lower scores on voice acceptability tests.

Nonspeech sounds. Other sounds—laughter, sighs, coughs, throat clearings, breath noises—also occur in voice communications and can provide information about the speaker's state or intentions. For naturalness, a voice transmission system should be able to transmit these sounds in a reasonably faithful manner. Low data rate voice algorithms that are optimized for speech may produce some rather odd effects when they encounter nonspeech sounds (Kemp, Sueda, and Tremain, 1989). Voice tests, seldom if ever, include such sounds, although they can influence acceptance of the system in actual use.

Contextual information. Context from several sources can help the listener to understand the spoken message. Grammatical constraints, realized in the sentence structure, provide one kind of context that helps us to know what kinds of words to expect next. In addition to the obvious constraints of English grammar in ordinary language, military language has its own special structures that constrain the word order and the type of words that are used to convey a message. Context is also provided by semantic constraints, that is the meaning of the words in a sentence or paragraph provides clues and expectations about what will be said next. The comparison of everyday sentences with semantically anomalous but grammatically correct sentences can be used to evaluate some of the effects of context.

When a voice system is used in the real world, situational knowledge is an important contextual factor. This can be anything from knowing the topic of a conversation to knowing what to expect at any given time in a fairly constrained scenario like air traffic control or ordering in a fast food restaurant. An important effect of context is that it limits the number of alternatives that can occur at any given time in the message. As the size of the response set decreases, intelligibility scores are higher and decrease more slowly with increased noise levels (Miller, Heise, and Lichten, 1951).

Speaker recognition. Although speaker identity is not directly related to intelligibility, speaker recognition can be an important aspect of user acceptance and deserves at least to be mentioned. Testing for speaker recognition can be very difficult and is not discussed in this report.

OVERVIEW OF SPEECH EVALUATION TECHNIQUES

Intelligibility tests evaluate the number of words or speech sounds that can be correctly identified in a controlled situation. The resposes can be objectively* scored as a percentage of correct responses. Acceptability or quality† tests evaluate the acceptability of the system based on listener judgments of subjective voice quality. In an attempt to avoid some of the problems of using human listeners, various physical measures of the speech signal have also been used with variable success to predict speech intelligibility or acceptability.

Intelligibility Test Methods

The basic methods of intelligibility testing for voice communication systems have been in existence for a long time. As early as 1910, Campbell used consonant-vowel (CV) syllables to test telephone transmissions. The classic paper by Fletcher and Steinberg (1929) describes test methods and results using a variety of speech materials, including consonant-vowel-consonant (CVC) nonsense syllables, multisyllable utterances, and English words, and it also included sentence comprehension based responses to queries, e.g., *Explain why a corked bottle floats*. However, the 1960 standard for monosyllabic intelligibility (ANSI 1960) contains the following statement.

> At present it is not possible to compare with precision two systems or conditions by testing one system or condition in one laboratory and the other system or condition in another laboratory.

Much subsequent research has been aimed at developing highly controlled and repeatable methodologies to reduce test-to-test variability, allowing for more accurate replicability and comparison of tests conducted at different times and places. In addition to the phonetically balanced (PB) monosyllabic word test specified in the 1960 standard, the current standard on speech intelligibility (ANSI 1989) includes two rhyme tests, the Diagnostic Rhyme Test (DRT) and the Modified Rhyme test (MRT). (The DRT and the MRT are independent tests; the MRT is a modification of an earlier test (Fairbanks, 1958).)

This section gives an overview of intelligibility test methods with selected examples of different test types. This review is not intended to be exhaustive or to cover all possible tests or types of tests that have been used to evaluate speech intelligibility. Rather, selected, commonly used, or promising test methods are described with a discussion of some of their major advantages and disadvantages. An excellent summary of a large number of different intelligibility tests can be found in Webster (1972); Kryter (1972) also discusses various aspects of intelligibility testing for speech communication. The three tests included in the ANSI (1989) standard are discussed first. These tests have been thoroughly validated and have been used extensively for testing voice systems, and there is a large body of literature that uses these tests. The standard describes methods for conducting intelligibility tests, including the selection and

*The terms *objective* meaures and *subjective* measures have been used in different ways in various contexts. As used in this chapter, objective refers to any measure that can be objectively scored, for example, the percentage of correct responses or the number of requests for repeats. Subjective refers to expressions of opinion (which may be assigned numeric values). Some authors have used the term subjective tests for all tests using human listeners, regardless of the basis for scoring, and the term objective to describe aspects of the speech signal that can be physically measured. I prefer to call the latter *physical* measures and to retain the distinction between objective and subjective aspects of listener behavior.

†Acceptability and quality have been used to refer to subjective judgment tests. The term acceptability is probably more accurate (Voiers, Panzer, and Sharpley, 1990), but the term quality is also in widespread use for such tests. To avoid confusion as to whether these are separate types of tests, both terms are used in this chapter.

training of the speakers and the listeners, how to conduct the test sessions, and the analysis of the results. The guidelines presented in the standard are applicable to other monosyllable and rhyme test materials as well as those described in the standard.

Standard Tests

The PB words, the DRT, and the MRT are all tests of phoneme intelligibility. In a phoneme intelligibility test, scores are based on the number of phonemes correctly identified by the listeners. Most frequently, single syllable words or nonsense syllables are used, and they may be spoken either as isolated utterances or in a carrier phrase, for example, "You will write the word _____ now." A phoneme test can be either open response, where the listener writes down the word or syllables that was heard, or closed response, where listener is given two or more choices in a multiple-choice format and selects the word that is closest to the one heard. A detailed discussion of the controls that need to be exercised over the test procedures, listener and speaker selection, recording and reproduction equipment, and either aspects can be found in the ANSI (1989) standard.

PB words. The PB word test is an open response test consisting of 1000 monosyllabic CVC words grouped into 20 lists of 50 words each, usually presented in a carrier phrase. Several variants of this type of test have been developed that are used for testing the hearing impaired. The Harvard PB word test (Egan, 1948) was, until recently, the only standard method for testing voice systems (ANSI, 1960), although closed response rhyme tests are now more common. Correctly training the speakers and listeners for this test is a cumbersome and expensive procedure often taking several weeks. Even then the listeners' scores continue to improve gradually with repeated testing, so that it is difficult to compare scores obtained at different times or in different laboratories.

Diagnostic Rhyme Test. The DRT (Voiers, 1977, 1983) is a two-alternative, closed response test that consists of 96 rhyming word pairs, in which the initial consonants differ only by a single distinctive feature (e.g., moot-boot differ only in the feature nasality). The features are derived from the Jacobson, Fant, and Halle (1952) distinctive feature system. The words are presented without a carrier phrase, so more words can be presented in the same amount of time than for tests using a carrier phrase. In addition to an overall intelligibility score, the DRT provides diagnostic feature scores on six phonemic features: voicing, nasality, sustention, sibilation, graveness, and compactness, and on a number of subfeatures, e.g., sibilation in voiced and unvoiced phonemes. A correction for guessing is used in scoring the DRT. Standard input tapes are available that use the same 6 speakers in the quiet and in a variety of military background noises such as helicopter, jeep, or tank. A test session usually uses 10 listeners, and the 2 who are the most inconsistent are eliminated, leaving 8 listeners for the final scores.

Modified Rhyme Test. The MRT (House, Williams, Hecker, and Kryter, 1965) is a closed response test that consists of 300 words presented as 6 lists of 50 words each and uses a 6-alternative format (e.g., rust, just, dust, must, gust, bust). A carrier sentence is usually used. Some of the alternatives differ from the target word by only a single phonemic feature and some differ by more than one feature. Both syllable initial and syllable final consonants are tested. A correction for guessing is not usually used in scoring.

Other Rhyme Tests. Rhyme tests have also been developed for other languages. Spelling modifications have been made to ensure that DRT pairs rhyme for British pronunciation (Pratt, Flindell, and Belyavin, 1987). A 6-alternative test exists that is similar to the MRT for German; diagnostic rhyme tests for French and for Dutch have also been developed following the same principles of the DRT (Peckels and Rossi, 1971; Steeneken, 1982). A PB monosyllabic word test also exists for Dutch (Houtgast and Steeneken, 1971). Finally, it should be mentioned that a "rhyme" test for vowels also exists (Clarke, 1965), which may be helpful in evaluating speech synthesizers.

Both the DRT and the MRT have been extensively used to evaluate military and commercial voice systems. This means that a large amount of historical data exists that can be compared with the results of any new tests. Scores on the DRT and MRT tend to be very highly correlated with one another over a wide variety of speech degradations (e.g., Voiers, 1983), so given the score on one of the tests, it is possible to predict the other and to make comparisons between two. The DRT has the advantage of providing diagnostic feature scores, and the MRT has the advantage of testing final as well as initial consonants. A diagnostic counterpart of the DRT, the Diagnostic Alliteration Test (DALT) (Voiers, 1981), has been developed for final consonants and is highly correlated with the DRT and MRT. Scores on the MRT tend to be lower for final consonants than for initial consonants, and scores on the DAT are also consistently lower than equivalent DRT scores (Voiers, 1981).

Closed response tests, such as rhyme tests, are easier to administer and score, and they produce more consistent results with less variability than open response tests. Test procedures for closed response rhyme tests can be standardized to the point where scores of the same system obtained at different times are highly repeatable. The Department of Defense (DoD) Digital Voice Processor Consortium, an interagency consortium to coordinate secure voice research, has conducted over the years numerous DRT tests of processors and modems. When voice systems were tested more than once, scores were usually within one or two percentage points of one another, with the exception of extremely low scores, which have greater variability. The average standard error for over 100 DRT scores listed in Sandy (1987) was 0.84. Closed response tests are also ideally suited for computerized data collection and scoring. The amount of practice needed to obtain stable performance is relatively small, and proper randomization procedures can be used effectively to prevent continued improvement.

Open response tests have the advantage that the listeners can indicate the sounds they actually heard and are not limited to the choices provided by the test developer. Another possible advantage is that scores on open response tests are usually lower, so with very good systems less possibility exists for *ceiling effects* (scores that are so near the maximum that differences in performance are indistinguishable). These advantages are offset by the greater variability in scores for open response tests, which means that there is less possibility for discriminating among closely competing systems. More listeners can be used to compensate for this effect. A more serious drawback is that practice and learning effects are also greater for open response tests, where the same pool of words is used repeatedly in different randomizations. Considerable practice is needed before performance becomes reasonably stable, and gradual continued improvement occurs even after extensive practice. This effect makes it very difficult to compare accurately the results obtained at different times or in different laboratories. Open response tests are also relatively expensive to administer and score because the written responses must be tallied by hand.

Sentence Tests

Tests that use complete sentences evaluate a number of speech cues not included in simple phoneme tests. Words in sentences tend to be less carefully articulated than words spoken in isolation, but sentences are often more intelligible because of grammatical and semantic constraints. Sentence structure also affects suprasegmental cues—pitch, intensity, and segmental duration. Sentence intelligibility is usually scored on the basis of the number of key words in the sentence that are correctly transcribed.

Harvard Sentences. The Harvard Sentences (Egan, 1948) consist of sets of phonetically balanced everyday sentences (e.g., *The birch canoe slid on the smooth planks*), scored on the basis of five key words in each sentence.

Haskins Sentences. More recently, a set of grammatically correct but semantically anomalous sentences (e.g., *The old corn cost the blood*), generally known as the Haskins Sentences (Nye and Gaitenby, 1973), has also been used by a number of researchers. The content words in each sentence are scored. These sentences serve to evaluate sentence comprehension with grammatical constraints but without semantic constraints to aid in word recognition. The Haskins and Harvard sentences are often used in the same experiment to separate the effects of grammatical and semantic constraints.

Like open response tests, sentence tests are cumbersome to administer and score. Sentence tests are by nature very difficult to adapt for use as a repeatable standardized intelligibility test, since they cannot be used repeatedly with the same listeners because once a sentence has been heard and understood, it is known to the listener. This means that a sustained testing program would require either an enormous supply of listeners or an inexhaustible source of sentences that have been equated for difficulty. However, sentence tests can be useful for one-time comparisons in a controlled experiment, when it is important to evaluate the intelligibility of connected speech, as for example with speech synthesis.

Other Speech Materials

A variety of other speech materials (e.g., polysyllabic words, paragraphs) and methods (memory, comprehension, reaction time) have been used to evaluate the effects of speech systems on performance. Many of these methods cannot be adapted to the repeated testing required for an extended testing program and are more suitable for limited experimental comparisons where only the speech systems tested at the same time can be meaningfully compared. Other types of material, for example consonants in other than initial and final positions (consonant clusters, intervocalic consonants), can be used to provide additional information in evaluating certain types of systems, e.g., synthesizers. Schmidt-Nielsen (1983) found that for linear predictive coded (LPC) speech at 2400 bits/s., the confusions for intervocalic consonants were different from the confusions for initial consonants as tested by the DRT. To be a useful evaluation tool, any test using a different type of speech materials would have to be extensively tested and validated, and the procedures would have to be standardized to assure that reliable results can be obtained that will be comparable across a variety of situations.

Experiments have also been conducted by testing specialized vocabularies such as military phrases (e.g., Webster 1972). The intelligibility of the International Civil Aviation Organization (ICAO) spelling alphabet and digits has been compared with DRT intelligibility for a number of digital and analog conditions (Schmidt-Nielsen, 1987a, 1987b). The spelling alphabet uses a small, highly distinctive vocabulary, so it is subject to ceiling effects for most ordinary voice systems, but it can be useful for evaluating the usability of very degraded systems where the DRT or other standard intelligibility scores are so low as to be considered unacceptable for normal use. The spelling alphabet is suitable for generating multiple randomizations and for standardizing procedures for repeated use. It can be readily adapted for machine scoring, as the letter or number responses can be typed on an ordinary keyboard. A randomization procedure has been developed that tests all of the letter names and digits while including some repetitions so that listeners will not develop expectations about the remaining words. Tape recordings have also been made using four of the speakers who have been used for DRT recordings.

Acceptability Test Methods

Acceptability or quality tests deal with subjective opinions of how the speech sounds. It is important to evaluate acceptability in addition to intelligibility because some degradations may affect one more than the other. Although subjective quality is often highly correlated with intelligibility, situations exist in which intelligibility may be high but speech quality is degraded. For example, a high-pitched whine in the background may not reduce intelligibility noticeably but could be so annoying as to make the speech almost intolerable to listen to. Likewise, certain degradations like peak clipping may have a relatively small effect on intelligibility but can still make the speech sound unpleasant (e.g., Licklider and Pollack, 1948). There can also be circumstances in which speech quality is improved but speech i·telligibility is still poor. Noise removal techniques, for example, can improve acceptability scores (Kang, 1989) but often lead to lower segmental intelligibility scores (Sandy and Parker, 1984).

The two most commonly used procedures for quality tests are pair comparisons and rating scales, or category judgments. Pair comparisons may be made among a set of voice systems of interest, pairing each system with every other system, or the system(s) of interest may be compared with a set of standard reference systems consisting of controlled noise levels or other distortions. The listener hears a sentence for each of two speech conditions and selects the one that is preferred. All pairs should be presented twice so that each system is presented both in first and second place. For ratings, listeners hear one or more sentences and are asked to rate the quality or acceptability of the speech on some form of rating scale. The listeners may be instructed either to assign labels—category judgments (e.g., excellent, good, fair, poor, bad)—which can later be converted to numerical scores, or they may assign numerical values directly. Reference systems are usually included in the tests to make comparisons with previous results easier.

The IEEE Recommended Practice for Speech Quality Measurements (1969) outlined three quality measurement techniques—the isopreference method, the relative preference method, and the category judgment method. The former two are pair comparison methods. The isopreference method developed by Munson and Karlin (1962) uses isopreference contours based on speech level and noise level. The relative preference method (Hecker and Williams, 1966) compares the test system to five reference systems consisting of different degradations. More recently, subjective quality has been referenced to varying amounts of multiplicative white noise, and the quality of the speech transmission device or distortion is given in terms of subjective speech-to-noise ratio (Nakatsui and Mermelstein, 1982). The selection of a single reference dimension such as subjective speech-to-noise levels allows for standardization of comparisons, but it is not at all clear that it is possible to make valid comparisons among widely different distortions (as in different kinds of digital voice algorithms) along a single dimension. The use of several different reference distortions raises the problem of selecting the appropriate distortions and also increases the number of pair comparisons that must be made. The use of pair comparisons tends to be very inefficient for comparing more than a very few systems, both in terms of listening time and in terms of the time needed to generate the test materials. If large numbers of reference systems are used, or if many different voice systems are to be compared directly with one another or with all of the reference systems, the number of pairs to be compared becomes very large. If several different speakers are used, the problem is multiplied.

The use of ratings or category judgments instead of pair comparisons greatly simplifies the data collection, since each system under test needs to be rated only once for each speaker. Experimental evidence (e.g., Voiers, 1977b) as well as many informal experiences in different speech laboratories indicate that the rank orderings assigned by the use of rating scales and by pair comparisons are very highly correlated. The most serious problem with using rating tests (and to some extent also pair comparisons) for speech quality is listener variability. Listeners can vary widely in their preferences, in how they use the scales, in the extent to which they spread their responses over the entire scale or use a portion of the scale. A variety of control procedures such as listener training listener normalization can be used to provide greater repeatability and stability of scores. Speaker variability can be controlled by using the same speakers in all tests, but caution should be used in comparing data for different speaker sets.

Diagnostic Acceptability Measure (DAM). The DAM (Voiers, 1977b; Voiers, Panzer, and Sharpley, 1990) is the quality test that has been taken the farthest toward standardizing test procedures, and developing techniques for reducing variability, and compensating for individual differences among individuals, and for changes over time. Voiers et al. (1990) list the measures used to control systematic and random error in the tests: (1) direct and indirect estimates of acceptability, (2) separate evaluation of signal and background quality, (3) explicitly identified anchors to give the listeners a frame of reference, (4) probes to detect shifts in listener adaptation level, (5) listener screening procedures, (6) listener training procedures, (7) listener calibration procedures, and (8) listener monitoring procedures.

The DAM uses standard input tapes consisting of 12 sentences for each of 6 speakers, 3 males and 3 females; listening crews include at least 12 listeners. There are 21 rating scales, 10 signal quality scales, 8 background quality scales, and 3 scales that evaluate overall speech characteristics. The rating scales are negatively oriented and evaluate the detectability of the effect in question on a 10-point scale that ranges from undetectable to overwhelming. The category labels (e.g., barely detectable, very conspicuous) for the ratings were experimentally selected from a large set of possible labels to be as nearly perceptually equidistant as possible. Potential listeners are screened for normal hearing, for their ability to discriminate different features of the speech signal, and for their consistency in using the rating scales. Upon selection and periodically thereafter, all listeners are calibrated against a large set of speech systems for which normative historical data have been obtained. This yields a set of constants and coefficients based on each individual listener's mean and standard deviation that can be used to transform that listener's scores to those of a theoretical "normative listener." These transformations are applied to future data obtained for that listener and are periodically updated. Anchors and probes are also built into each listening session, and additional adjustments are made based on probe performance for individuals who may be having a particularly lenient or strict day. The scales used on the DAM were derived from factor analysis. An overall composite acceptability score is arrived at by a complex set of combining equations. The diagnostic scales and the summary scores on the DAM have been validated against an extensive set of systematic speech degradations (Voiers, et al., 1990).

Over the years, the DoD Digital Voice Processor Consortium has conducted a large number of DAM tests, including many repetitions. In most cases, scores have been within one or two points of one another. On one occasion there was a difference of as much as five points after several years, but a source of bias was later identified in this case. The recently revised DAM II introduced some modifications to the way the questions are presented to the listener. This version should result in even better repeatability than with the previous one. Data collected over a period of 2 1/2 years (Voiers, et al. 1990) indicate that the standard deviation for inter-run variation of DAM scores was 1.01 points for unadjusted scores and 0.59 for probe adjusted scores.

Mean Opinion Score (MOS). The MOS refers to a general procedure that has been widely used for evaluating telephone systems (CCITT 1984a) and can have many variations. The speech material, typically sentences, is played through the voice system of interest and presented to listeners for scoring. The listener assigns scores on a five-point scale defined by category labels such as excellent, good, fair, poor, and bad. Sentences spoken by several different speakers may be used, often two males and two females. Large numbers of naive listeners rather than small numbers of trained listeners are generally used. A modification of the absolute category rating procedure, the Degradation MOS (DMOS) (CCITT 1984b), uses a five-point degradation or annoyance scale and includes a high-quality reference system to obtain greater sensitivity than the usual MOS procedure, especially for high-quality systems (Pascal and Combescure, 1988). Instead of calibrating the listeners, scores are referenced to the modulated noise reference unit (MNRU) (CCITT 1984c). Reference systems that are degraded by modulated noise at various speech-to-noise levels are included in tests of voice systems, and the scores for each system are referenced to equivalent Q-levels or modulated signal-to-noise (S/N) level. At least five reference systems at different S/N levels should be included. Goodman and Nash (1982) had tests of a number of communication and reference circuits conducted by using the same general procedure in seven different countries. They reported that average MOS scores differed considerably from country to country but that much of this variability was due to an additive scale shift. The average standard deviation for U.S. listeners was 0.75 on a scale of 1 to 5. There were 31 listeners, so the standard error in this case would be 0.135.

Phoneme specific sentences. A set of phoneme specific sentences was developed by Huggins and Nickerson (1985) for subjective evaluations of speech coders. Different sets of sentences contain consonants belonging only to certain phonetic categories or combinations of categories. For example, the sentence *Nanny may know my meaning* has only nasal consonants and vowels. The listeners first had to rank the speech conditions for each sentence and later provide degradation ratings for the same materials. Different types of sentence were sensitive to different aspects of degradation because of LPC processing. This type of test could provide a form of diagnosticity that is different from that provided by the DAM and perhaps more similar to that provided by the DRT. No listener calibration or standardization procedures were used.

Communicability Tests

Communicability tests are a variant of acceptability or quality tests that use two-way conversations. Standard quality and intelligibility tests using prerecorded materials are essentially one-way in that the speaker does not have the opportunity to modify his manner of speaking to fit the situation. In ordinary two-way conversations, there is feedback between the speaker and listener, and the person can speak more loudly or enunciate more clearly if necessary. Communicability tests use a two way communication task followed by a rating questionnaire, which results in subjective opinion scores not objective intelligibility scores. Communicability tests are cumbersome to administer because all of the voice systems to be tested need to be assembled and set up in the same location. So far, communicability tests have not been standardized for individual listener differences, so the scores are relative within a single test series and are not comparable across different tests. This means that all of the voice systems to be compared must be tested at the same time. Communicability tests are useful primarily for determining whether system deficiencies are compensable or noncompensable (Voiers and Clark, 1978). For example, a noisy connection can be overcome by speaking more loudly, but other degradations, such as the garbling caused by high bit errors in a digital transmission, may be more difficult to overcome.

Free Conversation Test. Conversational methods have been widely used in Britain to evaluate telecommunication systems (e.g., Richards and Swaffield, 1958; Richards, 1973). A task or problem to solve is given to each of two participants, who discuss the problem in a natural manner over the voice system to be evaluated and then rate the quality of the communication link after they are finished. Scores are Mean Opinion Scores based on a five-point rating scale of effort. In one version (Butler and Kiddle, 1969), each participant is given one of two pictures taken a short time apart, and they discuss the pictures until they can agree on which came first. Reference systems are generally included for comparison, but listener calibration procedures for communicability tests are not well developed. The test materials are not reusable in that once a given problem has been solved, the solution is known to the participants. The problems given to the participants with the Free Conversation Test also tend to vary in difficulty, and the time to reach a solution may vary from problem to problem.

Diagnostic Communicability Test. This test (Voiers and Clark, 1978) uses five participants and a stock trading game. The stocks assigned to each person vary from game to game; therefore the test materials are reusable and are also consistent in difficulty. There are 15 rating scales including both signal and background diagnostic scales. To implement this test, a five-way communication setup is needed to test the systems.

NRL Communicability Test. The NRL Communicability test (Schmidt-Nielsen and Everett, 1982) is a reusable variation on the Free Conversation method. The test uses an abbreviated version of the pencil and paper "battleship" game. Each player places two "ships" on a 5 by 5 grid, and the players take turns shooting at each other by specifying cells in the grid, e.g., alfa four or charlie three. The game can be used repeatedly with the same participants because the players place their own "ships" at the beginning of each game. The speech content is relatively uniform from game to game because the vocabulary used in playing the game is quite limited. There is some variability in game duration, but the time per move can be determined; however, Schmidt-Nielsen (1985) found that measures of speaker behavior, such as time per move, changes in vocabulary, or requests for repeats were less reliable measures of performance than were the subjective ratings. Scores obtained on the NRL test are relative and are not repeatable from one occasion to another, so reference systems should be included for comparison, and it is best to test all of the systems that need to be compared with one another at the same time.

Physical Measures of the Speech Signal

It is appealing to try to use physically measurable characteristics of the speech signal to evaluate intelligibility or voice quality because of the accuracy and repeatability of physical measurements compared with tests involving human listeners. Physical measures are also considerably cheaper and less time consuming than listener tests. Such measures can be useful for limited applications but should not be considered as substitutes for listener tests.

Articulation Index (AI). The AI (French and Steinberg, 1947) is computed as the average of the estimated articulation, based on speech to noise levels, in each of 20 contiguous frequency bands that contribute equally to speech intelligibility. The approximate relations of AI to the intelligibility of various types of speech materials including PB word sets of different sizes, nonsense syllables, rhyme tests, and sentences are given in Kryter (1972, p. 175). These are valid only for male speakers (p. 190). Plots of AI and PB word intelligibility for wideband and especially for narrowband noise (pp. 191, 192) show considerable scatter of the scores around the prediction curves, thus indicating that considerable discrepancy can exist between listener results and AI.

Speech Transmission Index (STI). The STI (Steeneken and Houtgast, 1980) is an improvement on the AI because it takes into account distortion from adjacent bands. The STI uses an artificial test signal and measures the effective signal-to-noise ratio in seven octave bands calculated from the modulation index of each band. The STI has been implemented in measuring devices—the STIDAS and RASTI (rapid STI). The STI has been shown to be very effective for noise and for auditorium measurements. Steeneken and Houtgast (1980) found a high correlation between STI and PB monosyllable intelligibility for Dutch for a variety of speech degradations and even for some wideband voice coders, and Anderson and Kalb (1987) also found a high correlation between STI and PB words for English. However, Schmidt-Nielsen (1987c) notes that the prediction errors (5.6%) associated with these correlations were too large for comparisons among systems with similar scores to be useful.

Combined Measures. As part of an ongoing program to develop predictors for speech acceptability, Barnwell and his colleagues (e.g., Quackenbush, Barnwell, and Clements, 1988; Barnwell, 1990) have tested a large number of objective measures of the speech signal as possible predictors of subjective speech quality. The test conditions covered a large number of different distortions including a variety of coding algorithms and a set of controlled distortions that included several levels of each of a variety of different distortions, such as additive noise, bandpass filtering, interruption, clipping, and voice coders. Subjective quality scores for each of the distortions were obtained using the DAM, and regression techniques were used to evaluate many possible objective measures, including S/N measures of various kinds, a number of different distance measures, and many others. No single measure performed very well in predicting acceptability over the entire database, although some were found to be very good for subsets of the database. Composite measures performed better than simple measures but even the best (based on 34 regression coefficients) had a correlation coefficient of only 0.84 and a standard error of estimate of 4.6. Segmental S/N ratio was an important attribute, and the frequency variant segmental signal to noise ratio was an excellent predictor for the subset of waveform coders, with a correlation of 0.93 and a standard error of estimate of 3.3 for this subset.

Physical measures of the speech signal can be a convenient method for estimating the effects of simple distortions such as noise, but it is important to realize the limitations of such measures for more complex distortions. They should not be used in making comparisons among different types of distortions or different classes of speech processing techniques. They are also not appropriate for evaluating non-waveform coders such as LPC systems. Listener tests are essential in evaluating the effects of voice processor improvements because coder distortions can interact in complex ways with perceptual processes.

RELATIONS AMONG DIFFERENT TESTS

Different tests are often highly correlated with one another because many of the degradations that occur caused by digital processing, background noise, or channel degradations affect many characteristics of the speech signal globally. Also, a number of degradations exist that affect some characteristics of the speech more than others, and in these cases one would expect tests that evaluate different characteristics to give divergent results. This section reviews some of the interrelations, similarities as well as discrepancies, that have been found in different types of speech materials and evaluation methods.

A large body of research suggests that although the difficulty may vary, measures of speech intelligibility are often highly intercorrelated. Fletcher and Steinberg (1929) showed systematic relationships between intelligibility scores for various sized speech units—phonemes, syllables, sentences—for a variety of telephone circuit conditions. Miller, Heise, and Lichten (1951) demonstrated a systematic effect of the size of the response set on the percentage of correct responses under varying degrees of noise degradation. Correlations have also been found between rhyme tests and other types of speech materials (Kryter and Whitman, 1965), including military vocabularies (Montague, 1960; Webster, 1972). A considerable number of the comparisons of different types of speech materials have involved systematically degrading the speech signal along a single dimension using different levels of the same type of degradation, often noise or bandpass limiting. With systematic degradations along a single dimension, one would expect tests using different speech materials to be highly correlated even though they might differ in difficulty. A high degree of cross predictability between the DRT and the MRT has been demonstrated by using a variety of different degradations (Voiers, 1983), but it should be noted that the speech materials for these two tests are very similar.

A discrepancy between the DRT and MRT has been noted in testing speech synthesizers. Pratt and Newton (1988) tested several speech synthesis systems by using the DRT, the MRT and another test, the Four Alternative Auditory Feature (FAAF) test. They obtained different rank orderings of the synthesizers with the DRT than with the other two tests, which gave results comparable to one another. It can be speculated that the discrepancy may be attributable to the fact that the DRT tests only initial consonants, while the MRT and FAAF test both initial and final consonants. Logan, Pisoni, and Greene (1985) found different groupings of synthesizers for final consonants than for initial consonants on the MRT. Unlike analysis-synthesis systems, synthesis-by-rule systems do not necessarily yield similar performance on initial and final consonants.

A number of researchers have found that different degradations can affect different kinds of speech materials in different ways. Hirsch, Reynolds, and Joseph (1954) compared nonsense syllables and one, two, and multisyllable words for different noise levels and for high- and low-pass filtering. The relationships among the different types of speech materials were not the same for the different degradations. Williams and Hecker (1968) used four different test methods (PB words, the Fairbanks Rhyme Test, the MRT, and lists of Harvard Sentences) to evaluate several different types of speech distortion—additive noise, peak clipping, and a channel vocoder at different error rates. They also found that the relationships among test scores and the rank orderings for the different speech distortions were not the same across speech materials, and they concluded that results for a given test were highly dependent on the nature of the distortion. Greenspan, Bennett, and Syrdal (1989) found very similar DRT scores for unprocessed speech and for two digital vocoders, but acceptability as measured by the DAM was considerably lower for both coders than for the unprocessed speech. When they used an open response consonant intelligibility test and naive listeners, both coders scored well below the unprocessed speech, and one of the coders had a lower score than the other. Tests with a larger number of response alternatives are generally more difficult than tests with a small number of alternatives and may be less subject to ceiling effects for high-intelligibility voice systems.

Schmidt-Nielsen (1987a, 1987b) conducted several tests comparing DRT intelligibility to the intelligibility of the ICAO spelling alphabet (alfa, bravo, charlie, etc.) and the digits zero to niner. Digital voice test conditions included LPC conditions with different levels of random bit errors and an 800 bit/s pattern matching algorithm. Analog conditions used speech transmitted over AM radio with noise jamming. These included continuous jamming conditions of varying degrees of severity and several interrupted jamming conditions. Within the LPC conditions, scores on both tests decreased as the bit error rate increased, and there was a consistent relationship between the two sets of scores. Both DRT and spelling alphabet scores also decreased with increased severity of the radio jamming conditions, but the relationship between the two tests was less consistent, especially for the interrupted conditions. The relationship between DRT scores and spelling alphabet scores was quite different for digital than it was for analog speech degradations. LPC conditions with low DRT scores showed very poor spelling alphabet recognition, but the noise-degraded, radio jamming conditions with similarly low DRT scores showed much higher spelling alphabet recognition. A DRT score near 50 corresponded to spelling alphabet intelligibility of just over 50% for LPC with bit errors but to spelling alphabet intelligibility of about 80% for noise jamming. This result makes sense in terms of the way the different degradations affect the speech materials on the two tests. DRT scores would be expected to be more vulnerable to noise degradation than are spelling alphabet scores. The DRT is based on consonant discriminations; consonants have less acoustic energy than vowels and can be expected to degrade considerably in noise. The spelling alphabet was developed specifically to be robust in noise, so the letter names differ from one another in their main vowels as well as in the number and pattern of the syllables. LPC, in contrast, is an analysis-synthesis system, and errors in the bit stream cause the wrong signal to be reconstructed at the receiver, which affects the entire speech signal, so that the vowel and prosody cues that help the spelling alphabet in the noise context are not as well preserved under the digital degradations.

It is not unusual to encounter speech systems for which the intelligibility is good but for which the voice quality is degraded on some dimension. However, when intelligibility is poor, judged voice quality usually goes down as well. An exception to this seems to occur when noise reduction techniques are used to remove or ameliorate background noise. The results of a National Research Council panel on noise removal (1989) indicate that noise reduction may lead to a subjective impression of improvement, but that there seems to be no evidence for any overall improvement in intelligibility as measured by standard phoneme intelligibility tests such as the DRT. Tests of a noise reduction preprocessor by the Digital Voice Processor Consortium (Sandy and Parker, 1984) using military background noises (such as helicopter or tank) indicated that intelligibility as measured by the DRT did not improve. In several cases

DRT scores were actually lower with noise reduction than without, while only one noise (helicopter) showed any improvement. In contrast, Kang and Fransen (1989) tested the same noise conditions using the DAM and found dramatic improvements in quality using a spectral subtraction technique for noise suppression. Speech samples with background noise were processed through a 2400 bit LPC voice processor with and without the noise processor as the front end. There was improvement in all cases, the average improvement was 6 points, and the greatest improvement was 13 points.

The purpose of testing is ultimately to determine the adequacy of the speech system for use in a real environment, although selection among several candidates for an application may also require the ability to make fine discriminations among closely competing systems. In actual use, the factors discussed in an earlier section that can affect intelligibility and acceptability interact in complex ways to determine how well the speech is understood and whether users find the system acceptable. In testing, the demands that are made on the listener vary with the type of task and the test materials that are used. The kind of information that is available to the listener to determine the correct response varies with the type of speech materials that are used. With rhyme tests, the listeners must rely almost entirely on segmental information for their responses, whereas with meaningful sentences, they have access to context from other words in the sentence and to information about grammatical structure from suprasegmental cues. The way in which the degradation interacts with the relevant speech cues should be considered in selecting test methods.

SELECTING TEST METHODS

Speech system evaluation is conducted in a variety of contexts with different goals and requirements, depending on the purpose of the test and the type of speech system to be evaluated.

Reasons for Testing. Some important reasons for speech evaluation tests might include developmental testing, diagnostic evaluation of defects, comparison and selection, operational usability evaluation, and the development of procurement specifications. Selection and specification testing rely heavily on standard test methods and highly controlled procedures that produce reliable numerical scores, whereas developmental and usability tests more often include nonstandard materials and evaluation experiments that cannot be generalized beyond the immediate context.

During the development of new voice systems or voice processing techniques, testing needs to be carried out regularly to monitor progress, to determine the weaknesses of the system, and to evaluate improvements. At times a very specific test may be needed to evaluate a particular aspect of the system that needed improvement, while at other times a wide variety of tests may be desirable to determine strengths and weaknesses and to guide future efforts. Much of this testing is highly informal, often consisting simply of listening to the output or perhaps asking one or two colleagues for an opinion. Periodically, more formal tests need to be carried out to monitor progress and to guard against the listener becoming so accustomed to the system that its defects are no longer noticed. Caution should be exercised in relying too heavily on a single standard intelligibility test when developing and refining new techniques. It is possible to tune the system too much to the particular features of a given test to the detriment of overall performance. A discrepancy between subjective quality evaluations and intelligibility scores can be an indication that this is happening.

The most common application of standard intelligibility and quality tests is for decision making or selection purposes. This may involve the selection of the best system for a particular application from several competing candidates or for comparing a newly developed or improved system with existing systems. Controlled test procedures that eliminate variability in the test results due to irrelevant factors are required to make fair comparisons. It is highly desirable to simulate conditions that may occur in use such as environmental background noise or channel degradations.

Tests of user acceptance of telephone devices may include user opinions of test sentences or conversations based on ratings or customer interviews. Operational evaluation of military voice systems usually involves field tests in which the developers of the systems are not involved. When tests are conducted in operational environments, it is often difficult to get the users to conduct controlled conversations or to be analytic about the quality of the voice system. If the user does not like it or if some part of the system fails to operate properly when needed, the system is unacceptable, even if the reason for the failure is unrelated to the quality of the voice transmission. Laboratory tests may be conducted for the purpose of predicting the usability of a device in a given environment. Estimates of user acceptance may also be developed from experiments establishing correlations of standard tests with field tests or user evaluations.

When establishing specifications of minimally acceptable intelligibility levels for procurement contracts, very exact test procedures are needed. If the specification establishes a target intelligibility score that must be met or exceeded, the test must be capable of producing scores that are repeatable from one occasion to the next with very little variability. With tests that produce good discriminations but have less numeric stability, one alternative is to specify that the test score must be within a specified number of points of the score of a reference or standard system to be tested at the same time.

Type of Voice Application. The type of speech system to be evaluated and the nature of the degradation of the speech signal should be considered in relation to their effect on the factors that influence speech intelligibility and acceptability. The way in which the test materials and tasks affect and interact with the relevant types of speech information, such as segmental cues, prosody, and context, can then be used to select the tests that will be the most informative. If background noise or poor channel conditions are likely to be present in the intended application, they should be included in the test program, and the robustness of the competing systems under these conditions would be an important consideration in the selection process.

For voice communication systems that start with human speech at the input and reproduce a more or less faithful version of the same speech as output at the receiving end, a consonant test is a reasonably accurate predictor of overall intelligibility. In this type of application, it can be assumed that if the consonant sounds are reproduced correctly, the other sounds will also be good. Existing wideband algorithms for digital telephone transmission generally have very good to excellent intelligibility. Sometimes the intelligibility may be high enough to produce ceiling effects on standard rhyme tests, giving little discrimination among the scores for competing methods. Where intelligibility is so high as to be considered good enough for normal communications, minor differences may be unimportant, and the quality of the voice system becomes the overwhelming consideration. For most narrowband systems, it is likely that some loss in speech intelligibility and quality may occur, so ceiling effects are less of a problem. Although some modern narrowband systems approach wideband systems in intelligibility, very low data rate systems can have substantially reduced intelligibility. Some military applications include adverse conditions where intelligibility may fall into ranges that would be unacceptable for normal communications but that may still be quite usable for the restricted and distinctive vocabularies used in many military communications.

No single test is equally sensitive to small differences across the entire range of intelligibility. Different speech materials and test formats vary in the difficulty of the acoustic discriminations needed for the responses. In general, we would expect a difficult test to be more sensitive to small amounts of degradation, but when the speech quality is very poor, the test would lose sensitivity because of floor effects. A test with easier discriminations or more context would discriminate well among poor speech conditions but would be subject to ceiling effects for less degraded speech.

Both intelligibility and quality tests are important for evaluating voice transmission systems. The choice between the DRT and the MRT depends to some extent on the historical background. If a particular application has used the MRT extensively in the past, it should continue to be used, even though the DRT provides more detailed diagnostic information. The DoD Digital Voice Processor Consortium has used the DRT since the early 1970s and has accumulated a large historical database of DRT scores for wideband and narrowband digital voice systems for military applications (e.g., Sandy and Parker, 1984; Sandy, 1987). The MOS has been widely used in telephone applications, but the DAM produces highly repeatable results, offers diagnostic scales, and has been extensively used to evaluate digital voice systems for military applications (Sandy and Parker, 1984; Sandy, 1987). A disadvantage of the DAM is that the details of the scoring procedure are at present proprietary to a single company that provides testing services. Many of the advantages of the DAM could be reproduced by using appropriate diagnostic scales and rigorous listener calibration and monitoring procedures with other test materials. Communicability tests can provide two-way conversations under controlled conditions at considerably less expense than field tests. For systems or conditions where intelligibility and quality can be expected to be very poor, it may be useful to conduct supplementary tests. Research at NRL with the ICAO spelling alphabet indicates that a test using this limited and distinctive vocabulary can be useful for evaluating usability when DRT scores fall into the unacceptable range.

For other applications, such as synthesis-by-rule systems, it is necessary to consider other phoneme categories and word positions. Rhyme tests can give an indication of synthesizer perfoımance (Pratt, 1987), but the speech materials on such tests are too limited for a complete evaluation of synthesis systems. The MRT tests both initial and final consonants, but a combination of the DRT for initial consonants and the DALT for final consonants would give more diagnostic information about specific weaknesses. It is important also to include tests of consonants in other positions, such as word medial position or consonant clusters, as well as vowels in different consonant contexts and at different stress levels. Intonation and timing rules that follow the constraints of the spoken language must also be explicitly built into the system, and the effects of prosody on intelligibility and naturalness need to be evaluated. Tests using sentence materials can be used to evaluate multiple phonemic contexts as well as the effects of prosody on intelligibility and quality. The Harvard sentences and Haskins sentences may be useful, but they do not provide the possibility of repeated use necessary for making accurate comparisons of different synthesis systems. A subjective acceptability test that provides numeric stability and diagnostic scales would also be very useful in evaluating the naturalness of speech synthesizers.

In general, the more highly processed the speech signal (i.e., synthesis, very low data rate speech algorithms, noise removal techniques), the more important it is to include tests that evaluate several different attributes of the speech signal. The same is true when comparing different speech processing methods or different types of speech degradations.

GENERAL RECOMMENDATIONS

Although the choice of test method and the extent of testing depends on the purpose of the tests and requirements of the users, it is essential to use careful controls and proper procedures to ensure that the test results are valid. Refer to the ANSI (1989) standard method for measuring the intelligibility of speech over communication systems for details of experimental control for conducting intelligibility tests.

1. *Whenever possible, use standard test materials.* It is highly desirable to use standard evaluation methods, so that comparisons can be meaningfully made among different types of voice systems and among tests conducted at different times and in different places. If standard tests are inappropriate for the application or if additional tests are needed, preference should be given to tests or materials that have been used by other researchers and for which historical data are available. If reliability data are available, tests with high reliability are preferable.

2. *Always include reference conditions.* It is extremely important to include reference systems, such as high quality unprocessed speech and several known degradations for which previous historical data are available, to provide a context for interpreting the scores. This is especially important if nonstandard speech materials, unknown speakers, or untrained listeners are used. When rhyme tests like the DRT and MRT are used with standard speakers and scored by laboratories with trained and screened listening crews, known reference conditions are often available and need not be included with every new test.

3. *Use multiple speakers.* Given that speaker differences can be quite large, at least six to twelve speakers should be used for most intelligibility tests. A larger number of speakers may be needed for sentence materials than for rhyme test materials. Male and female speakers should be used unless the application is known to be restricted to only one sex.

4. *Use a sufficient number of listeners.* The IEEE Recommended Practice for Speech Quality Measurements (1969) recommends 6 to 10 trained listeners or at least 50 untrained listeners. These numbers, or a few more listeners are also reasonable for intelligibility tests. The DRT procedure starts with 10 listeners and eliminates the 2 that are the most inconsistent over an entire session, so that there are actually 8 listeners for each score.

5. *Exercise the system.* For communication systems, test different environmental conditions that may occur in use, e.g., background noise, channel degradations. For synthesizers, include a variety of speech materials to test phonemes in different contexts as well as sentence materials to evaluate prosody. In general, the more highly processed the speech is, the more important it is to evaluate several different types of speech materials covering different types of speech cues that contribute to intelligibility.

6. *Compare.* When comparing very different processing methods or speech degradations, use several different types of tests and speech materials. When comparing similar processing methods or degradations, a more limited set of speech materials may be used, but it may be useful to include a greater variety of environmental conditions.

7. *Exercise meticulous care in recording and playing tapes.* Use high quality equipment and follow correct procedures for the selection, setup, and maintenance of recording and playback equipment and the storage and copying of tapes. Seemingly minor deficiencies like dirty tape heads or a bad connector can significantly reduce speech quality and render the outcome of an entire test series invalid.

8. *Use proper statistical procedures to compare test scores.* The measurement error inherent in using human listeners has been discussed in various sections of this report. Proper statistical methods are needed to make the correct comparisons among the different systems that have been tested. These might include, for example, t-tests, analysis of variance, and multiple comparison tests. When in doubt, consult someone with a broad knowledge of behavioral statistics.

ACKNOWLEDGMENTS

I would like to thank the many colleagues and friends who have read and commented on an earlier version of this report. I am especially grateful to Thomas H. Crystal, John D. Tardelli, and William D. Voiers for their thoughtful reading and helpful suggestions. I apologize for the long delay in the publication of this report, which was also intended as a chapter in a book (publication date still undetermined). Unfortunately, this means that some recent results are not included here.

REFERENCES

American National Standards Institute. (1960). *American Standard Method for Measurement of Monosyllabic Word Intelligibility* (ANSI S3.2-1960, American Standards Association, New York, N.Y.)

American National Standards Institute. (1989). *Method for Measuring the Intelligibility of Speech Over Communication Systems* (ANSI S3.2-1989 - A Revision of ANSI S3.2-1960, American Standards Association, New York, N.Y.)

Anderson and Kalb. (1987). "English Verification of the STI Method for Estimating Speech Intelligibility of a Communications Channel." *J. Acoust. Soc. Am.* **81**, 1982-1985.

Barnwell, T.P. (1990). "A New Objective Speech Quality Measure for Speech Coding Systems," *J. Acoust. Soc. Am.* **87**, S13.

Butler, L.W. and L. Kiddle. (1969). "The Rating of Delta Sigma Modulating Systems with Constant Errors and Tandem Links in a Free Conversation Test Using the Reference Speech Link," Report No. 69014, Signals Research and Development Establishment, Ministry of Technology, Christchurch, Hants.

Campbell, G.A. (1910). "Telephonic Intelligibility," *Phil. Mag.* Jan. 1910.

CCITT. (1984a). "Absolute Category Rating (ACR) Method for Subjective Testing of Digital Processors," *Red Book* V, (Annex A to Suppl. 14).

CCITT. (1984b). "Subjective Performance Assessment of Digital Encoders Using the Degradation Category Rating (DCR) Procedure," *Red Book* V. (Annex B to Suppl. 14).

CCITT. (1984c). Recommendation P.70 (Subjective voice-ear measurements - modulated noise reference unit) *Red Book* V, 111-114.

Clarke, F.R. (1965). "Technique for Evaluation of Speech Systems," Contract DA 28-043 AMC-00227(E), Final report of Stanford Research Institute Project 5090 on U.S. Army Electronics Laboratory.

Egan, J.P. (1948). "Articulation Testing Methods," *Laryngoscope* **58**, 995-991.

Fairbanks, G. (1958). "Test of Phonemic Differentiation: The rhyme test," *J. Acoust. Soc. Am.* **30**, 596-600.

Fletcher, H. and J.C. Steinberg. (1929). "Articulation Testing Methods," *Bell System Tech. J.* **8**, 806-854.

French, N.R. and J.C. Steinberg. (1947). "Factors Governing the Intelligibility of Speech Sounds," *J. Acoust. Soc. Am.* **19**, 90-119.

Goodman, D.J. and R.D. Nach. (1982). "Subjective Quality of the Same Speech Transmission Conditions in Seven Different Countries," *IEEE Trans. Commun*, COM-30, 642-654.

Greenspan, S.L., R.W. Bennett, and A.K. Syrdal. (1989). "A Study of Two Standard Intelligibility Measures," *J. Acoust. Soc. Am.* **85**, S43.

Hecker, M.H. and C.E. Williams. (1966). "Choice of Reference Conditions for Speech Preference Tests," *J. Acoust. Soc. Am.* **39**(5, Pt. 1), 946-952.

Hirsh, I.J., G. Reynolds, and M. Joseph. (1954). "Intelligibility of Different Speech Materials," *J. Acoust. Soc. Am.* **26**(4), 530-538.

House, A.S., C.E. Williams, M.H. L. Hecker, and K.D. Kryter. (1965). "Articulation Testing Methods: Consonantal Differentiation With a Closed Response Set," *J. Acoust. Soc. Am.* **37**, 158-166.

Houtgast, T. and H.J.M. Steeneken. (1971). "Evaluation of Speech Transmission Channels Using Artificial Signals," *Acoustica* **25**, 355-367.

Huggins, A.W.F. and R.S. Nicherson. (1985). "Speech Quality Evaluation Using 'Phonemic-Specific' Sentences," *J. Acoust. Soc. Am.* **77**, 1896-1906.

IEEE Subcommittee on Subjective Measurements. (1969). "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Trans. Audio and Electroacoustics* **17**, 227-246.

Jacobson, R., C.G.M. Fant, and M. Halle. (1952). "Preliminaries to Speech Analysis: the Distinctive Features and their Correlates," Tech Rep. No. 13 Acoustics Laboratory, MIT.

Kang, G.S. and L.J. Fransen. (1989). "Quality Improvement of LPC-Processed Noisy Speech by Using Spectral Subtraction," *IEEE Trans. Acoust. Speech and Signal Processing* **ASSP-37**, 939-942.

Kemp, D.P., R.A. Sueda, and T.E. Tremain. (1989). "An Evaluation of 4800 BPS Voice Coders," *IEEE ICASSP-89.*

Kryter, K.D. (1972). Speech Communication. In H.P. Van Cott and R.G. Kincade, Eds., *Human Engineering Guide to Equipment Design* (U.S. GPO, Washington, DC).

Kryter, K.D. and E.C. Whitman. (1965). "Some Comparisons Between Rhyme and PB Word Intelligibility Tests," *J. Acoust. Soc. Am.* **37**, 1146.

Licklider, J.C. and I. Pollack. (1948). "Effects of Differentiation, Integration, and Infinite Peak Clipping Upon the Intelligibility of Speech," *J. Acoust. Soc. Am.* **20**, 42-51.

Logan, J.S., D.B. Pisoni, and B.G. Greene. (1985). "Measuring the Segmental Intelligibility of Synthetic Speech: Results from Eight Text-to-Speech Systems," Research of Speech Perception Progress Report No. 11, Indiana University, Bloomington, IN.

Luce, P.A., T.C. Feustel, and D.B. Pisoni. (1983). "Capacity Demands in Short-Term Memory for Synthetic and Natural Speech," *Human Factors* **25**, 17-32.

Miller, G.A., G.A. Heise, and W. Lichten. (1951). "The Intelligibility of Speech as a Function of the Context of the Test Materials," *J. Experimental Psychology* **41**, 329-355.

Montague, W.E. (1960). "A Comparison of Five Intelligibility Tests for Voice Communication Systems," Report No. 977, U.S. Navy Electronics Laboratory, San Diego, CA.

Munson, W.A. and J.E. Karlin. (1962). "Isopreference Method for Evaluating Speech Transmission Circuits," *J. Acoust. Soc. Am.* **34**, 762-774.

Nakatsui, M. and P. Mermelstein. (1982). "Subjective Speech-to-Noise Ratio as a Measure of Speech Quality for Digital Waveform Coders," *J. Acoust. Soc. Am.* **72**, 1136-1144.

National Research Council Panel on Noise Removal. (1989). "Removal of Noise from Noise-Degraded Speech Signals," Panel on removal of noise from a speech/noise signal, National Academy Press, Washington, DC.

Nye, P.W. and J. Gaitenby. (1973). "Consonant Intelligibility in Synthetic Speech and in a Natural Control (Modified Rhyme Test Results)," Haskins Laboratories Status Report on Speech Research, SR-33, 77-91.

Pascal, D. and P. Combescure. (1988). "Evaluation de la Qualité de la Transmission Vocale (Evaluation of the Quality of Voice Transmission)," *L'Echo des RECHERCHES* 132(2), 31-40.

Peckels, J.P. and M. Rossi. (1971). (Cited in Voiers, 1983). "Le Test de Diagnostic per Paires Minmales Adaptation au Francais du Diagnostic Rhyme Test de W.C. Voiers," *Journee d'Etudes sur la Parole,* Groupement des Acousticiens de Langue Francais, April, 1971.

Pollack, I. and J.M. Pickett. (1964). "Intelligibility of Excerpts from Fluent Speech: Auditory vs. Structural Context," *J. Verbal Learning and Behavior* 3, 79-84.

Pratt, R.L. (1987). "Qualifying the Performance of Text-to-Speech Synthesizers," *Speech Technology,* March/April, 54-64.

Pratt, R.L., I.H. Flindell, and A.J. Belyavin. (1987). "Assessing the Intelligibility and Acceptability of Voice Communication Systems," Report No. 87003, Malvern, Worcestershire: Royal Signals and Radar Establishment.

Pratt, R.L. and J.P. Newton. (1988). "Quantifying Text-to-Speech Synthesizer Performance: An Investigation of the Consistency of Three Speech Intelligibility Tests," Proceedings of Speech 1988, 7th FASE Symposium, Edinburgh.

Quackenbush, S.R., T.P. Barnwell, and M.A. Clements. (1988). *Objective Measures of Speech Quality* (Prentice Hall Press, Englewood Cliffs).

Richards, D.L. (1973). General background. *Telecommunications By Speech* (Butterworth and Co. (Publishers Ltd., London, England) p. 1-27.

Richards, D.L. and J. Swaffield. (1958). "Assessm[illegible] of Speech Communication Links," *The Institution of Electrical Engineers,* Paper No. 2605 R, 7[illegible]2.

Sandy, G.F. (1987). *Digital Voice Processor [illegible]nsortium Report on Performance of 16 kbps Voice Processors MRT-87W161* (Mitre Corp., McLean, VA).

Sandy, G.F. and Parker. (1984). *Digital Voice Processor Consortium Final Report MTR-84W00053* (Mitre Corp., McLean, VA).

Schmid-Nielsen, A. (1983). "Intelligibility of VCV Segments Excised from Connected Speech," *J. Acoust. Soc. Am.* **74**, 726-738.

Schmidt-Nielsen, A. (1985). "Problems in Evaluating the Real-World Usability of Digital Voice Systems," *Behavior Research Methods, Instruments, and Computers* **17**, 226-234.

Schmidt-Nielsen, A. (1987a). "Evaluating Degraded Speech; Intelligibility Tests Are Not All Alike," Official Proceedings of Military Speech Tech 1987 (Media Dimensions, Inc., New York, N.Y.) pp. 188-121.

Schmidt-Nielsen, A. (1987b). "The Effect of Narrowband Digital Processing and Bit Error Rate on the Intelligibility of ICAO Spelling Alphabet Words," *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-35**, 1101-1115.

Schmidt-Nielsen, A. (1987c). "Comments on the Use of Physical Measures to Assess Speech Intelligibility," *J. Acoust. Soc. Am.* **81**, 1985-1987.

Schmidt-Nielsen, A. and S.S. Everett. (1982). "A Conversational Test for Comparing Voice Systems Using Working Two-Way Communication Links," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-30**, 853-863.

Schmidt-Nielsen, A., H.J. Kallman, and C. Meijer. (1990). "Dual Task Performance Using Degraded Speech in a Sentence Verification Task," *Bulletin of the Psychonomic Society* **28**, 7-10.

Steeneken, H.J.M. (1982). (Cited in Voiers, 1983). "Ontwikkeling en Toetsing van een Nederlandstalige Diagnistische Rijmtest voor het Testen van Spraak-Kommunikatiekanalen," Report No. IZF, 1982-13. Institut voor Zintuigfysiologie TNO Soesterberg.

Steeneken, H.J.M. and T. Houtgast. (1980). "A physical Method for Measuring Speech-Transmission Quality," *J. Acoust. Soc. Am.* **67**, 318-326.

Summers, W.V., D.B. Pisoni, R.H. Bernacki, R.I. Pedlow, and M.A. Stokes. (1988). "Effects of Noise on Speech Production: Acoustical and Perceptual Analyses," *J. Acoust. Soc. Am.* **84**, 917-928.

Voiers, W.D. (1977a). "Diagnostic Evaluation of Speech Intelligibility," *In* M.E. Hawley, Ed. *Speech Intelligibility and Speaker Recognition* (Dowden, Hutchinson, and Ross, Stroudsburg, PA).

Voiers, W.D. (1977b). "Diagnostic Acceptability Measure for Speech Communication Systems, ICASSP-77," *IEEE International Conference on Acoustics, Speech, Signal Processing,* New York.

Voiers, W.D. and M.H. Clark. (1978). "Exploratory Research on the Feasibility of a Practical and Realistic Test of Speech Communicability," Final Rep. on Contract No. N0039-77-C-0111, Dept. of the Navy, Navy Electronics Systems Command, Washington, DC (Dynastat, Inc., Austin, TX).

Voiers, W.D. (1981). "Uses, Limitations, and Interrelations of Present-Day Intelligibility Tests," Proceedings of the National Electronics Conference **35**, Chicago, IL.

Voiers, W.D. (1982). "Some Thoughts on the Standardization of Psychological Measures of Speech Intelligibility and Quality," Proceedings of the Workshop on Standardization for Speech I/O Technology, National Bureau of Standards, Gaithersburg, MD.

Voiers, W.D. (1983). Evaluating Processed Speech Using the Diagnostic Rhyme Test, *Speech Technology*, Jan/Feb., 30-39.

Voiers, W.D., I.L.Panzer, and A.D. Sharpley. (1990). "Validation of the Diagnostic Acceptability Measure (DAM II-B)," Contract No. MDA904-87-C-6026 for National Security Agency, Ft. Meade, MD (Dynastat, Inc., Austin, TX).

Webster, J.C. (1972). Compendium of Speech Testing Material and Typical Noise Spectra for Use in Evaluating Communications Equipment (Technical Document 191). Naval Electronics Laboratory Center, Human Factors Technology Division, San Diego, CA.

Williams, C.E. and M.H. Hecker. (1968). "Relation Between Intelligibility Scores for Four Test Methods and Three Types of Speech Distortion," *J. Acoust. Soc. Am.* **44**, 1002-1006.